MDCN: Multi-Scale, Deep Inception Convolutional Neural Networks for Efficient Object Detection

Wenchi Ma[1]; Yuanwei Wu[1]; Zongbo Wang[2]; Guanghui Wang[1]
[1]University of Kansas, [2]Ainstein Inc.

Motivation

Object detection in challenging situations such as ***scale variation***, ***occlusion***, and ***truncation*** depends not only on **feature details** but also on **contextual information**.

- Previous: emphasize much on detail features by deeper and wider network
- Problem: low effectiveness of feature usage with high load of computation as feature details are easily being changed or even "washed out" after passing through complicated filtering structures.
- **MDCN: proposes multi-scale and deep inception convolutional neural network, focusing on wider and broader object regions by activating feature maps produced in deep part of the network.**

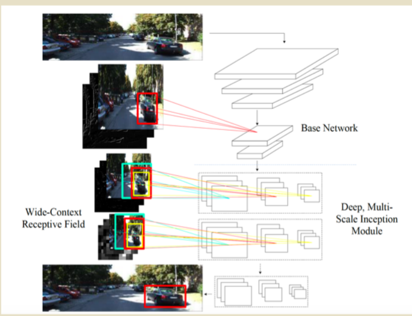


Fig.1 Multi-scale, wide-context receptive field activation

Contributions

- **Integrate the contextual information into the self-learning process** through a **single-shot** network structure.
- **Information square inception modules** are proposed to detect objects with **multi-size context expression** while maintaining a high computation efficiency by parameter sharing.
- The proposed MDCN model achieves better performance with a relatively shallow network at a real-time speed.

Detection Pipeline

Feature extraction, wide-angle contextual information, object classification and bounding box regression are performed in a single-shot pipeline.

1. Base network: VGG-16
- extract high-resolution, low-dimensional features
2. Multi-scale deep inception module:
- extract object main-body and multi-scale contextual information.

Wide-Context Receptive Field

Guide the network to activate various contextual regions by a spontaneous learning process.

- more sensitive towards main-body features of objects
- pay more attention to relationships among objects and between objects and scenes

Feature maps produced in deep layers cover larger proportion of the original scene

- Receptive fields are able to cover larger scope of scenes
- Various contextual information can be involved in actual learning course

$$\Phi_n = f_n(\Phi_{n-1}) = f_n(f_{n-1}(\ldots f_1(I)))$$

$$\Phi_m = F_m(\Phi_{m-1}) = F_m(F_{m-1}(\ldots F_{m-k}(\Phi_n))),\ m-k>n$$

$$F_j = f_i(\Phi_{j-1}; W_j),\ m-k \le j \le m$$

Information-Square Inception Modules

- Combination of 1x1, 3x3 and 5x5 filters: activating multi-scale receptive fields
- using two series of 3x3 filters to replace 5x5 filter so as to minimize the number of parameters

By defining weights to each filtering units, the information-square inception modules formed.

$$F_j = f_j(f_j(\Phi_{j-1})) + 2\times f_j(\Phi_{j-1}) + \Phi_{j-1},\ m-k \le j \le m$$

$$F_j^2(\Phi_{j-1}) = (f_j^2 + 2\times f_j + 1)(\Phi_{j-1}) = \left(\left(f_j+1\right)^2\right)\left(\Phi_{j-1}\right),\ m-k \le j \le m$$

Data and Implementation

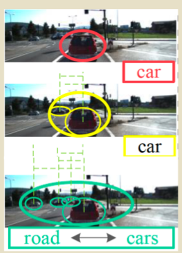
Dataset: KITTI

- containing many challenging objects like small and occluded cars, pedestrians and cyclists
- objects are labeled as easy, moderate, and hard based on how much objects are occluded and truncated

Implementation:

- All images are rescaled from 1242x375 to 300x300
- Intersection over Union (IoU) for car, pedestrian and cyclist are all set to 50%
- The VGG-16 base network is pretrained on ImageNet and MDCN is fine-tuned on KITTI

Detection Accuracy



object instance

object vs. object

object vs. scene

TABLE I
THE COMPARISON RESULTS OF DIFFERENT MODELS IN TERMS OF AVERAGE PRECISION(%) ON KITTI VALIDATION SET.

Model	Car			Pedestrian			Cyclist			mAP
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	
SSD	85.00	74.00	67.00	53.00	50.00	48.00	46.00	52.00	51.00	58
ResNet-101	87.57	76.04	68.07	50.27	47.74	45.21	49.86	53.61	51.77	58.9
WRN-16-4	90.08	76.8	68.5	52.29	47.88	45.3	47.71	50.36	49.38	58.7
WR-Inception	87.1	77.2	68.81	55.98	**52.51**	48.61	52.9	54.63	52.87	61.18
WR-Inception-12	**90.36**	78.24	71.11	53.26	51.08	**49.54**	57.02	59.28	57.39	63.03
MDCN-I1	88.40	87.96	87.34	**56.39**	50.37	48.86	71.58	72.21	**76.82**	71.91
MDCN-I2	88.70	**88.19**	**87.91**	55.02	50.21	48.28	**73.85**	**72.66**	74.95	**72.30**

TABLE II
RESULTS ON KITTI VALIDATION SET FOR DIFFERENT IOU THRESHOLDS.

Classes	Methods	IOU						
		0.5	0.55	0.6	0.65	0.7	0.75	0.8
Car	SSD	83.9	80.9	77.6	74.5	67.7	59.4	49.7
	MDCN-I1	88.1	87.4	84.3	79.0	**75.9**	**69.3**	59.6
	MDCN-I2	**88.4**	**87.6**	**85.2**	**79.1**	**75.9**	69.0	**59.7**
Pedestrian	SSD	47.3	41.2	32.7	27.3	20.8	15.9	**12.4**
	MDCN-I1	**54.8**	**48.4**	41.1	32.2	24.5	**18.8**	11.8
	MDCN-I2	54.0	47.4	**42.1**	**35.5**	**26.3**	15.9	9.7
Cyclist	SSD	61.5	52.0	48.7	41.0	30.2	21.7	11.0
	MDCN-I1	72.8	62.6	56.9	51.0	**41.0**	28.5	18.1
	MDCN-I2	**75.0**	**68.9**	**64.3**	**52.6**	40.1	**28.7**	**21.8**

TABLE III
DETECTION EFFICIENCY OF DIFFERENT MODELS

Model	Network	GPU	Resolution	# of Params	FPS
SSD	VGG-16	K40	300×300	2.41×10^7	17.0
MDCN-I1	VGG-16	K40	300×300	2.54×10^7	15.8
MDCN-I2	VGG-16	K40	300×300	2.55×10^7	15.4

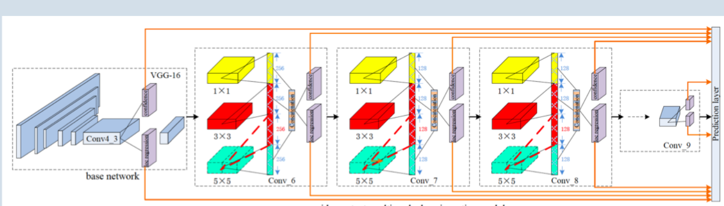


Fig.2 The architecture of MDCN. The wide-context, multi-scale deep inception module consists of multiple filtering structures. The red, yellow and green boxes each indicate one filter size.